SRA File Transfer Guide

National Center for Biotechnology Information (NCBI)
National Library of Medicine
Version 2.0 Draft G 30F7 January 2009

Contents

Notice

Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government, and shall not be used for advertising or product endorsement purposes.

1 Overview

This application note identifies needs and use cases for high throughput file transfer to the NCBI Short Read Archive (SRA). For such transfers, file sizes of 12 GB and larger are commonplace, and a single submission may involve hundreds of such files. As the sizes of the datasets have increased, we have found that the traditional methods of *ftp* or *http* do not have the performance characteristics needed to support this load of data.

Requirements for large scale data transfer over the internet include high bandwidth, auto checksum, recursive copy, and security based on strong keys. NCBI and EBI have chosen to use products from Aspera, Inc to exchange short read data with each other because of these improved data transfer characteristics, and instructions are provided below for submitters who wish to use the same mechanism. NCBI is open to using additional products with the appropriate performance characteristics as well, and continues to be committed to providing all users with needed data, whether by fasp, ftp, or hard disk drive (HDD) and tape.

1.1 Scope

This document is intended for users doing large data file transfers with NCBI. This document attempts to meet the particular needs of sequencing Centers and Brokers submitting short read data for the project.

1.2 Related Documents

Short Read Archive Submission Guidelines

1.3 Links and Contacts

http://www.ncbi.nlm.nih.gov/Traces/sra	SRA Home Page at NCBI
trace@ncbi.nlm.nih.gov	The Trace Archives mailing list for inquiries
Trace Help Desk	Specific problems or issues regarding trace submissions (web form)
sra@ncbi.nlm.nih.gov	Aspera technical support
www.asperasoft.com	Asperasoft home
http://en.wikipedia.org/wiki/EXT3	ext3 filesystem
http://en.wikipedia.org/wiki/NTFS	NTFS filesystem

http://en.wikipedia.org/wiki/Linear_Tape-Open	Linear Tape Open (LTO) tape formats
http://en.wikipedia.org/wiki/Ftp	File transfer protocol (ftp)
http://en.wikipedia.org/wiki/Http	Hypertext transfer protocol (http)
http://en.wikipedia.org/wiki/USB	Universal Serial Bus (USB) connections
http://en.wikipedia.org/wiki/User_Datagram_Protocol	User Datagram Protocol (UDP)

1.4 Revision History

2.0 Draft F: 27 January 2009 (shumwaym)	Reviewed and updated with ssh key pairs, dedicated accounts, new Aspera versions
2.0 Draft G: 30 January 2009 (roachtg)	Added '-T' ascp disclaimer to section 4.5

2 ftp

The ftp service provided to established centers has long been the normal method for transferring trace data with NCBI.   Users are recommended to switch to Aspera client for downloads, and to use *ascp* copy program for uploads.

2.1 Limitations using ftp

Traditionally NCBI has relied on *ftp* as the means for transferring large files. Bandwidth on transfers is typically 100 Mbps with less on international transfers. NCBI does not impose an upper limit on ftp transfer size. However, maximum file sizes above 10 GB may fail due to limitations elsewhere in the path from center to NCBI.

2.2 Bulk Submissions via ftp

High-volume submissions should be uploaded to the ftp directory for your center, which is provided with the secure ftp account provisioned to your center.

For example, a user working for the *mycentre* center will deposit short read data into the short read directory of the ftp account's login directory as follows:

```
ftp ftp-trace.ncbi.nlm.nih.gov
login: mycentre_trc
passwd: !jXYZZ3@ce

  > cd short_read
  > put myfiles.tar.gz
  > quit
```

You should double check that the file size that you posted agrees with the original file.

2.3 Individual Submissions via ftp

The NCBI Trace Archives maintains a private ftp address. Please write to mailto: sra@ncbi.nlm.nih.gov for the current address, which contains both the ftp address and login string. The unix/linux shell command will look something like:

```
ftp  ftp://sra:0!8e5frRy!@ftp-private.ncbi.nih.gov/
```

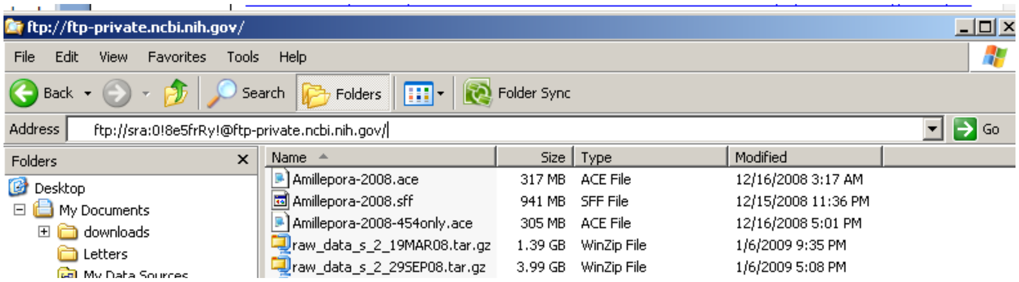
Please observe the following rules when using this submission method:

1. Maximum 1 Gigabyte file size
2. Maximum 10 file limit per submission
3. Choose a unique filename that also will be easy for you to identify

This directory has special access rules. You can stat the directory (list the files), but you cannot read any file (or download a file), and Once deposited, the file cannot be overwritten. The files are removed as soon as processed, or if they have remained too long on the server. It is your responsibility to complete the submission transaction in a reasonable amount of time so that the files you have deposited through this channel can be processed by the submission system.

2.4 ftp from Windows

It is possible to upload to NCBI ftp sites from Windows. Use Windows Explorer to access the individual ftp address as follows. Then simply drag and drop the submission files from your source directory into the destination directory that the Explorer tool has opened.



Address: ftp://sra:0!8e5frRy!@ftp-private.ncbi.nih.gov/

Name	Size	Type	Modified
Amillepora-2008.ace	317 MB	ACE File	12/16/2008 3:17 AM
Amillepora-2008.sff	941 MB	SFF File	12/15/2008 11:36 PM
Amillepora-2008-454only.ace	305 MB	ACE File	12/16/2008 5:01 PM
raw_data_s_2_19MAR08.tar.gz	1.39 GB	WinZip File	1/6/2009 9:35 PM
raw_data_s_2_29SEP08.tar.gz	3.99 GB	WinZip File	1/6/2009 5:08 PM

You can also login using your center account, and utilize Windows Explorer to navigate and upload.

2.5 Troubleshooting ftp

If you are having trouble with your ftp connection to NCBI, try

- Setting passive mode rather than active mode
- Ask your sysadmin to increase ftp buffer size to 32 MB
- Try another host, or another platform (Windows instead of Unix)

- Try using the unix *split* utility to split up the transfer file into smaller pieces. Be sure to provide reassembly instructions and checksum of the reassembled file to verify rebuilding the original file on our end.
- Try another ftp client software:
  - *ncftp (*http://www.NcFTP.com*)*
  - Windows *filezilla* (http://filezilla.sourceforge.net)

If you still have trouble, please write us with the following details:

- Time of transfer (GMT or local time)
- IP address of ftp client (the system you are doing ftp from)
- Version of unix software (*uname –a*, or *cat /proc/version*)
- ftp account used
- specific error messages (connection closed, etc)

3 Disk and Tape

Archive users can also request or submit data on disk or tape. The following are requested:

- LTO4 (we can also read LTO3 and LTO2)
- HDD with USB2.0 or FireWire interface enclosure with WinNT (FAT32) partition type, so any Windows or Linux computer can read them.
- NTFS, Ext3, or other large format drives. Please ensure they are delivered with an enclosure. We prefer FireWire interface.

To get the submitted media returned, please plan on providing a waybill for shipping. If you are requesting a download by disk or tape, please send us the media first, along with a waybill for return shipping.

Please use the following shipping address:

Martin Shumway, Staff Scientist
DHHS/NIH/NLM/NCBI
45 Center Drive
Bldg. 45/Room 6AS37D-53
MSC 6510
Bethesda, MD 20892
shumwaym@ncbi.nlm.nih.gov
tel: 301.402.4041
fax: 301.402.9651

4 Aspera

4.1 The fasp Protocol

The FASP protocol from Aspera (www.asperasoft.com) uses UDP, eliminating the latency issues seen with TCP, and provides bandwidth up to 1 Gbps to transfer data. It has a restart capability if data transfer is interrupted midstream and is well behaved, so if there is other data traffic on your network connections, it will back off in order to avoid starving other protocols. We have seen effective throughput up to 600 Mbps to a single site.

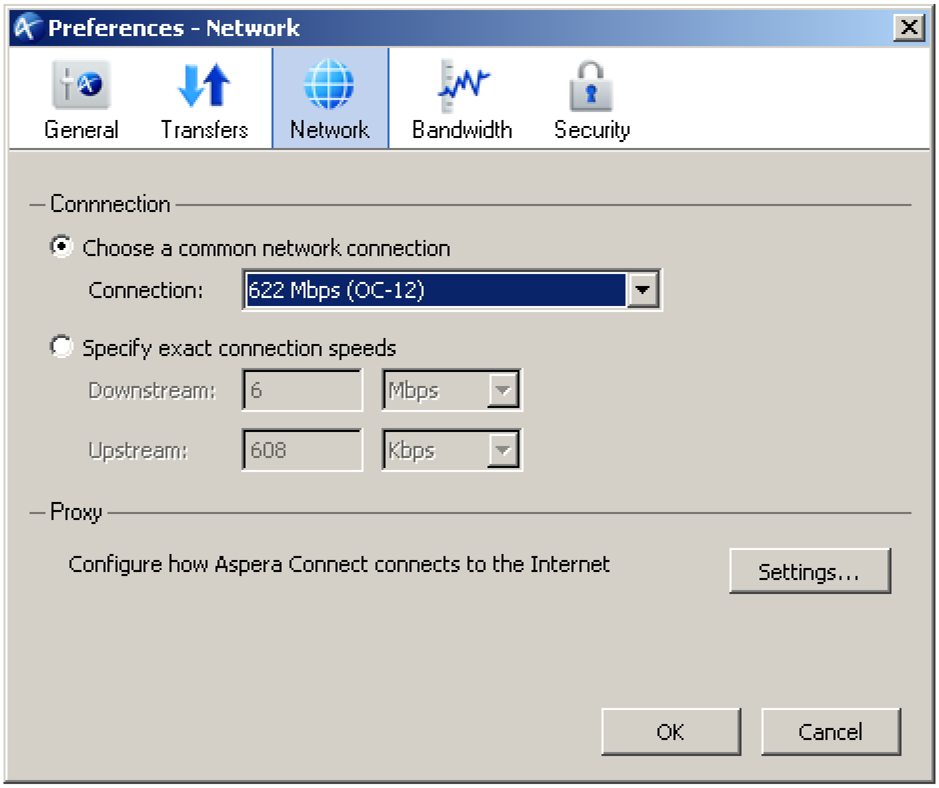
NCBI is implementing Aspera for two use cases, occasional users and those who download files for direct use (Aspera Connect), and bulk users who will be uploading or downloading large amounts of data (ascp)

4.2 Aspera Connect

Aspera Connect is software that allows download and upload via a web plugin for popular browsers for machines running Linux, Windows, and Macs and a command line tool that allows scripted data transfer. The software client is free for NCBI site users for the purpose of exchanging data with NCBI.

Download and install AsperaConnect software:
http://www.asperasoft.com/downloads

Select the Connect product for your browser and platform. By default, the plugin configuration is less than optimal. To change, right click the Aspera icon from the system tray. Select ‘networks’ and update the connection speed (e.g. 622Mbps).



Connection

● Choose a common network connection

Connection: 622 Mbps (OC-12)

○ Specify exact connection speeds

Downstream: 6 Mbps

Upstream: 608 Kbps

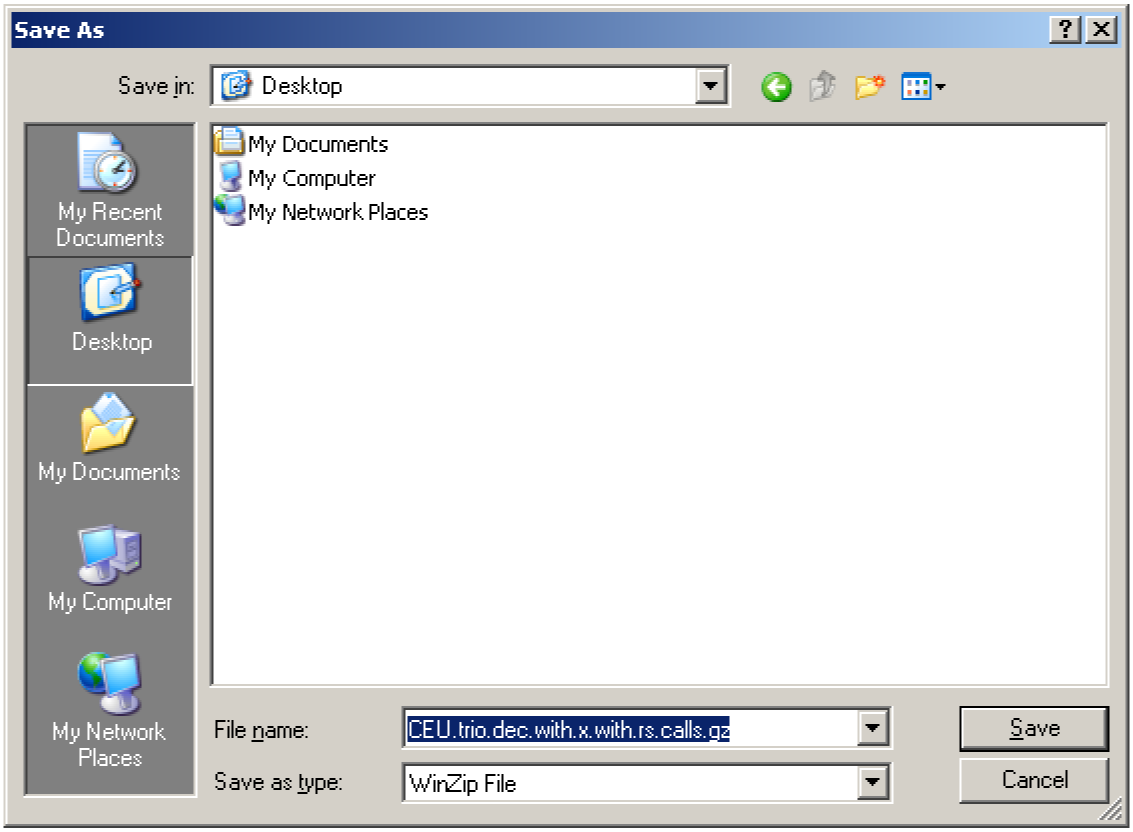
Proxy

Configure how Aspera Connect connects to the Internet

4.3 Pulling Data with Aspera Connect (from browser)

Note – Sometimes Mozilla requires that all Aspera connect files and two subdirectories (from Aspera\Aspera Connect\lib) be copied to the Mozilla FireFox plugins directory.

Once plugin has been installed in your browser, you may download file(s)/directory(ies) from NCBI using Aspera. For example, in your browser window, go to
http://www.ncbi.nlm.nih.gov/projects/faspftp/1000genomes/
Select a file (left click mouse).
Click “Save” to begin saving the data. You may be prompted to select where the file is to be saved. For example:



File name: CEU.trio.dec.with.x.with.rs.calls.gz

Save as type: WinZip File

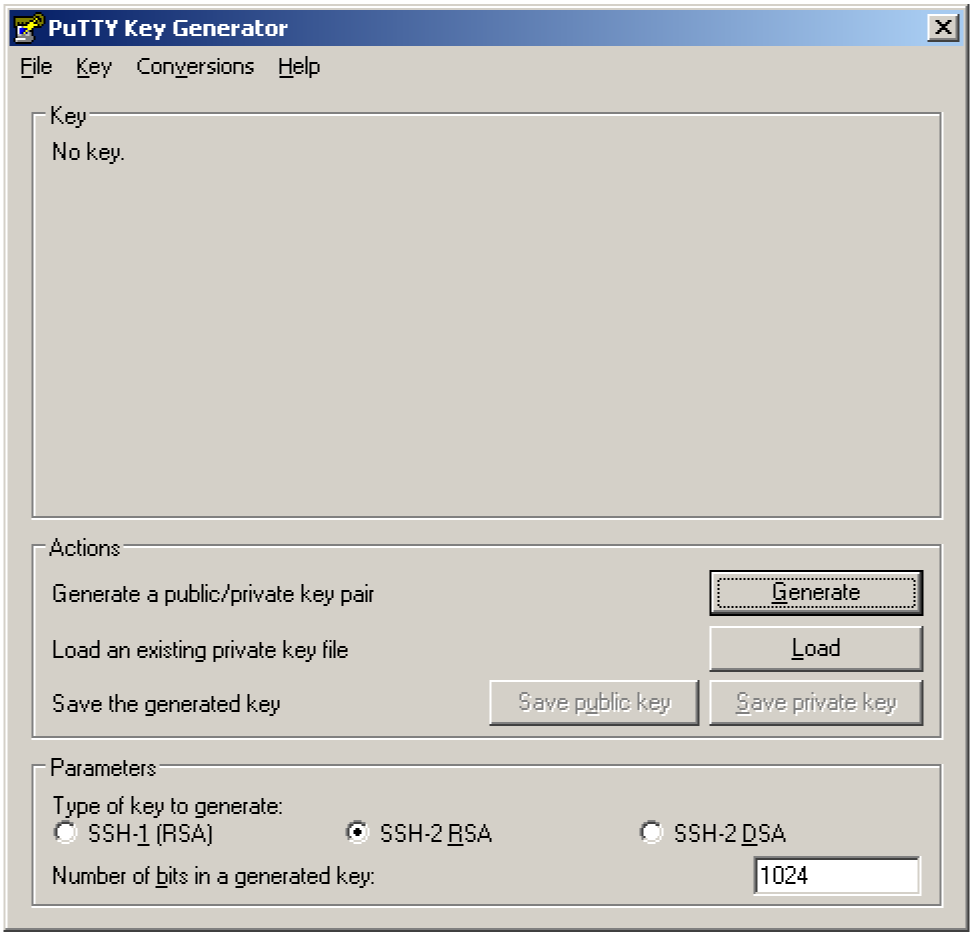
4.4 Setting Up Aspera for Bulk Transfers

Tell NCBI you are preparing to set up a link, and we will provide a login account. There will be one account per center. Please set up a Center identity for your institution or lab if you do not already have one.

Your local firewall must permit UDP data transfer on port 33001 in both directions to allow the fasp traffic to pass and to allow ssh traffic outbound to NCBI.

Download puttygen: http://the.earth.li/~sgtatham/putty/latest/x86/puttygen.exe

Run puttygen.exe to create ssh key:

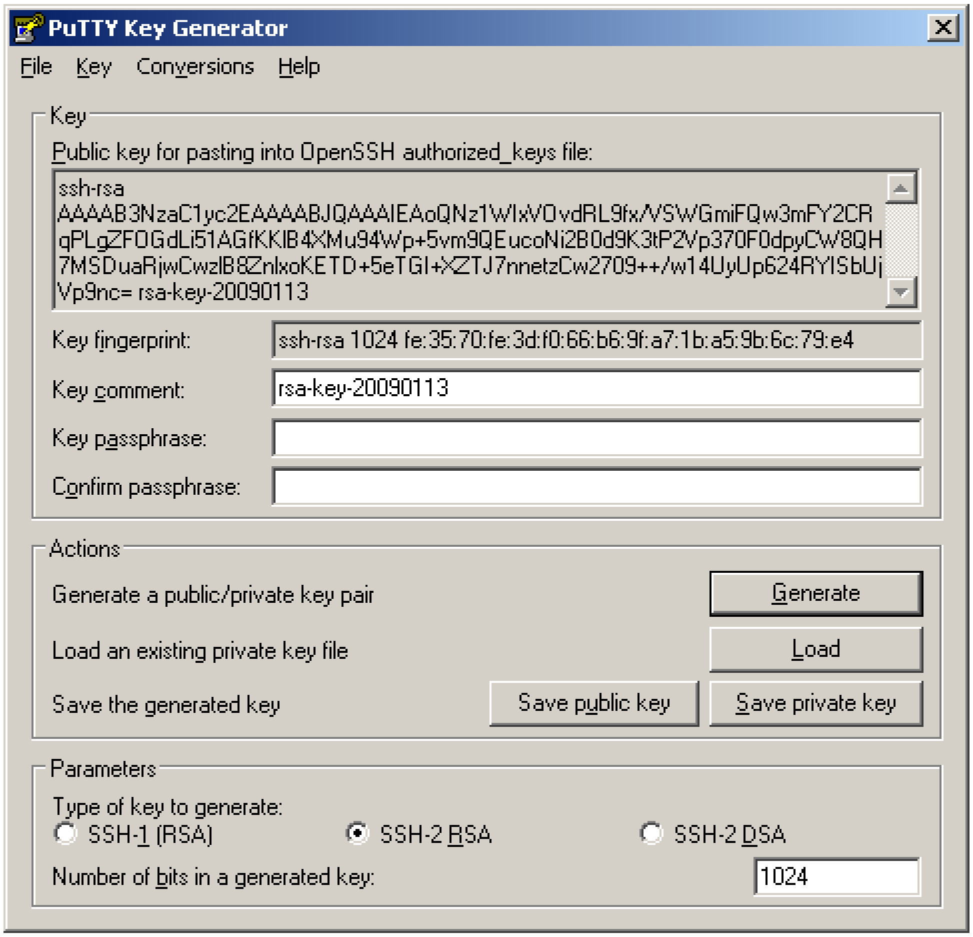


Type of key to generate:
○ SSH-1 (RSA) ● SSH-2 RSA ○ SSH-2 DSA

Number of bits in a generated key: 1024

Make sure that SSH-2 RSA Parameter option is selected, and that the "Number of bits in a generated key" be set to 1024. Then press "Generate" (moving mouse to generate key).

Generating a key will result in something like this:



Public key for pasting into OpenSSH authorized_keys file:

```
ssh-rsa
AAAAB3NzaC1yc2EAAAABJQAAAIEAoQNz1WIxVOvdRL9fx/VSWGmiFQw3mFY2CR
qPLgZFOGdLi51AGfKKlB4XMu94Wp+5vm9QEucoNi2B0d9K3tP2Vp370F0dpyCW8QH
7MSDuaRjwCwzlB8ZnlxoKETD+5eTGI+XZTJ7nnetzCw2709++/w14UyUp624RYISbUj
Vp9nc= rsa-key-20090113
```

Key fingerprint: ssh-rsa 1024 fe:35:70:fe:3d:f0:66:b6:9f:a7:1b:a5:9b:6c:79:e4

Key comment: rsa-key-20090113

Key passphrase:

Confirm passphrase:

Generate a public/private key pair

Load an existing private key file

Save the generated key

Type of key to generate:
○ SSH-1 (RSA) ● SSH-2 RSA ○ SSH-2 DSA

Number of bits in a generated key: 1024

Click “Save Private Key” to retain the private key. NOTE – leave “Key passphrase” and “Confirm passphrase” empty (otherwise, you will be prompted to enter the passphrase whenever you do an Aspera transaction).

Copy the text from the “Public Key for pasting into OpenSSH authorized_keys file” text box:

ssh-rsa
AAAAB3NzaC1yc2EAAAABJQAAAIEAoQNz1WIxVOvdRL9fx/VSWGmiFQw3mFY2CRqPLgZFOGdLi51AGfKKlB4XMu94Wp+5vm9
QEucoNi2B0d9K3tP2Vp370F0dpyCW8QH7MSDuaRjwCwzlB8ZnlxoKETD+5eTGI+XZTJ7nnetzCw2709++/w14UyUp624RYISbUjV
p9nc= rsa-key-20090113

In order that a submission center is able to access (i.e. transfer and receive files from NCBI using Aspera Connect), this public ssh key must be provided to NCBI. This key should be emailed to:
sra@ncbi.nlm.nih.gov with subject line “Aspera connect authorization request”.
SSH keys are used for establishing secure connections to remote computers

4.5 Using ascp for Bulk Transfers

The command line program *ascp* is a utility delivered along with the AsperaConnect product.

- You can run the *ascp* program with the following parameter settings:
  - –Q (for adaptive flow control)
  - –l (maximum bandwidth of request, try 200M and go up from there)
  - –m (minimum bandwidth of request, try 0)
  - –r recursive copy
  - –T no encryption (speeds up transfers). Connection remains secure however data being transferred is not.
  - –i <private key file>

- Try experimental transfers starting at 100 Mbps and working up to 400-500 Mbps. Select the bandwidth setting that gives good performance with unattended operation. Copy the file to:

  -
    ```
    ascp -i <private key file>  -QTr <file(s) to transfer>
    -l100M  asp-<center>@fasp.ncbi.nlm.nih.gov:test/
    ```

where

<private key file> ::= fully qualified path & file name where the generated private key was saved.
<files(s) to transfer> ::= names of files to transfer (including path)
<center> ::= name assigned to the submission center, provided by sra@ncbi.nlm.nih.gov if not already in existence.
100M ::= tunable mbit/sec bandwidth

- The *ascp* command on Microsoft Windows is located by default in `c:\program files\aspera\Aspera Connect\bin\ascp`
- The *ascp* program on Mac in located at *aspera/bin/ascp*
- The *ascp* program on Linux is located at *<install directory>/bin/ascp*
- It is possible to run *ascp* in an autonomous, unattended manner that does not require repeated login. Please send us the public key of a SSH key pair and we will add it to our authentication system.

4.6 Pushing Data with ascp

Use the command line utility *ascp* to copy files directly to a remote host:

```
ascp -i <private key file>  -QTr <file(s) to transfer>  -l300M  \
asp-<center>@fasp.ncbi.nlm.nih.gov:incoming/
```

where

<private key file> ::= fully qualified path & file name where the generated private key was saved.
<files(s) to transfer> ::= names of files to transfer (including path)
<center> ::= name assigned to the submission center, provided by sra@ncbi.nlm.nih.gov if not already in existence.
300M ::= tunable mbit/sec bandwidth

4.7 Pulling Data with ascp

Use the command line utility *ascp* to copy files directly from a remote host:

```
 ascp -i <private key file> -QTr –l300M <mysource> <mydestination>
```

where
<private key file> ::= fully qualified path & file name where the generated private key was saved.
<mysource> ::= fully qualified internet file name (host:filename)
<mydestination> ::= local directory
300M ::= tunable mbit/sec bandwidth

4.8 Administering Remote Files

Do **not** delete files in order to "make space". The SRA is responsible for maintaining adequate space by removing files that have already been processed. If files are not being deleted it is likely because of a backlog in the SRA.

If a submission file needs to be replaced, wait until you have a replacement and then overwrite the file (do not delete it). Please DO replace zero length files or files that have been truncated. If a "junk" file has been transmitted by mistake, it can be removed.

NOTE - files that have not been attached to any submission may be deleted after a certain amount of time. It is recommended that you consult with the SRA Administrators for the current expiration policy.

You may establish a secure connection to the SRA by using putty.exe along with your private ssh key. For example:

```
  putty.exe –i <private key file> asp-<center>@fasp.ncbi.nlm.nih.gov
```

where

<private key file> ::= fully qualified path & file name where the generated private key was saved.

<center> ::= name assigned to the submission center, provided by sra@ncbi.nlm.nih.gov if not already in existence

Once connected, you may use the '*ls*' command to view the directory. You will not be able to change directories (e.g., use of the '*cd*' command is disabled). Valid *ls* commands include:

```
ls -l test       #lists the content oftest subdir in long format
ls -l incomimg   #lists the content of incoming subdir in long format
ls -l            #lists the content of home directory in long format
ls -lR           #lists the content of all entries in home directory
```

To remove a file, use the *rm* command. For example:

rm incoming/badfile

4.9 Debugging ascp Transfers

- To make a test downloads using *ascp* please try this command:

  ```
  o  ascp -i <private key file>  -QTr <file(s) to transfer>
     -l100M  asp-<center>@fasp.ncbi.nlm.nih.gov:test/
  ```

  where

  <private key file> ::= fully qualified path & file name where the generated private key was saved.
  <files(s) to transfer> ::= names of files to transfer (including path)
  <center> ::= name assigned to the submission center, provided by sra@ncbi.nlm.nih.gov if not already in existence.
  100M ::= tunable mbit/sec bandwidth

- Be sure that the local storage is fast enough to sustain this rate. We have seen problems with download if the target storage is on slow network volume. If you wish, examine unix */var/log/messages* for a fasp log file, and send that to Aspera support.
- Note that when a submitter uses a wild card for submissions, 0 length files matching the shell expansion are created in the destination directory. These placeholders can be present for a time before the actual download takes place. Therefore, some buffer time should be added to any process on the transmission side that is responsible for determining whether the transfer succeeded.
- A connection error like this one may be due to expiration of license key, or incorrect private key:

  ```
  ascp: session open failed.
  >> ascp: (remote) failed to initiate session, consult log.
  ```

```
>> Ssh error: SSH connection failure: 130.14.29.99:22 Server
reported
>> failure exit code 1
```

4.10 Caveats

1. Supplying a directory as a source will cause the creation of the corresponding sub-directory tree on the destination. To avoid this, ensure that you execute the *ascp* command while in the source directory and provide a list of files to be transferred.

4.11 Known Problems

1. Please be aware that ':' (colon) character is not allowed in filenames by *ascp* command and files need to be renamed prior to transfer.